THE VARIABILITY OF TEACHERS' MARKS

NATHAN SILBERSTEIN
Stuyvesant High School, New York City

There have been many studies during the past few years that indicate that teachers do not mark alike, that there is a wide variation in the estimates of the values of answers to questions, expounded by pupils. Whether or not this variability is desirable depends largely upon whether or not it is possible to secure uniformity in teachers' personalities. If it is true that no two persons are alike, then it follows naturally that they will differ in their judgments and estimates. We, as individuals, are not estimated equally by our friends and acquaintances. But there are many elements in common in the judgments of every one of our "judges." We are not called "grouch" by one and "wit" by another; "generous to a fault" by one, and "the meanest man on earth" by another. If it were possible to create a character scale, it is highly improbable that any one individual would be rated at the two extremes of the scale by two different judges.

All teachers are familiar with this very phenomenon in rating examination papers. Very frequently one teacher will accept as perfect an answer that some other teacher rejected as worthless. This is particularly true in English, where so much depends upon the impression made upon the examiner. There is no exact standard such as one has in mathematics or in the sciences—although these, too, show very wide variations. However, it is with the question of variability in rating English that this discussion is concerned.

Last September, the Board of Regents of the state of New York rejected as unsatisfactory a number of papers in the fourth-year English written by students in the Stuyvesant High School in June, 1920. The number was greater than the chairman of the department, Dr. Frederick H. Law, thought allowable. Dr. Law felt that this was due to a lack of standardization of acceptable answers.

In order to arrive at a basis of discussion a particular paper, one written by a conscientious, faithful student, was selected. This paper had been rejected by the Regents although rated 73 per cent by the members of the English department, marking it in committee. Typewritten copies of this paper were distributed among all the members of the department to be rated by them as a whole. All identification marks, such as the pupil's name and class were removed, so that a disinterested judgment was possible. Thirty-one teachers rated the seven questions on the paper. These results were tabulated and an analysis attempted by the writer, who is not a teacher of English. The original manuscript that had been written by the pupil was also placed in his hands.

As most teachers who have had experience in marking Regents' papers know, each teacher initials his or her rating. It is thus possible to identify a particular rating. It soon became evident that there would be great variation in the department as a whole, for not one teacher of the seven who rated it originally, defended his original rating. For example, teacher A rated the first question 13 out of 15 on the original paper in June; on the typewritten copy his rating was 10. I have prepared a table showing the original ratings, the Regents' rating, the second ratings given by those teachers who had rated it originally, the maximum ratings allowed for each question, the median and the average rating for the seven questions asked.

TABLE I

	Questions							Total
	I	II	III	IV	V	VI	VII	
Original rating	13	20	6	6	8	15	5	73
Regents' rating	10	18	5	3	6	13	4	59
Second rating	10	17	5	5	7	17	4	65
Maximum rating	15	25	10	12	10	20	8	100
Median rating	11	18	5	7	6	10	3	60
Average rating	11	19	4.8	7.7	6.3	11.3	3.6	63.7

The lowest rating given the paper as a whole by any one teacher was 43, the highest 75. In other words one teacher thought the paper worth almost twice as much as did another teacher, both of whom were presumably expert in their fields.

It is, however, on the ratings of individual questions that the emphasis of this study is placed; for it was here that the lack of uniformity displayed itself. In the first question the pupil was asked to write an interesting letter of at least 150 words to a friend on "My favorite hobby or study." Fifteen counts were allowed. The boy wrote a letter telling why he liked mathematics. Table II and Figure 1 give the judgment of the 31 teachers:

TABLE II

Rating	No. of Teachers	Rating	No. of Teachers
7	2	12	7
9	1	13	5
10	9	14	0
11	6	15	1

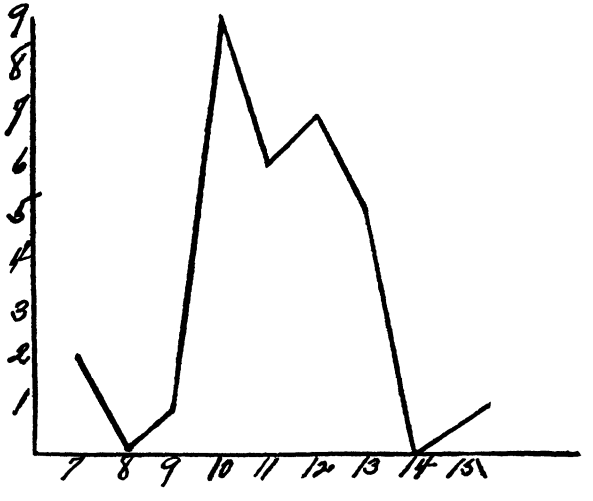

FIG. 1

This table reads as follows: two teachers rated this letter 7 out of 15, one rated it 9 out of 15, nine rated it 10, etc. One teacher gave it the maximum rating 15, calling it a perfect letter, yet two did not think it was even half-perfect. It seems that teachers of English ought to decide what standards of judging should be used in estimating the value of letter-writing in order to avoid doing the pupil an injustice. Combining the facts of Table II with those of Table I, the combined judgment of the teachers would

tend to indicate that the letter was probably worth from 10 to 12. The middle 50 per cent of the ratings fell between these values. It might be said that a teacher who rated it above 12 was too lenient and one who rated it below 10 too severe.

The second question was: "Discuss one of the following topics, using at least 200 words. Give substantial reasons for your statements: Events leading to the Revolutionary, Civil, or world-war. (Maximum credit 25.)" The pupil discussed the events leading to the Revolutionary War. Table III indicates the variability of the teachers' judgments.

TABLE III

Rating	No. of Teachers	Rating	No. of Teachers
10	1	19	2
12	1	20	7
15	4	21	1
17	3	22	4
18	7	23	1

It will be noted that thirteen of the thirty-one teachers—nearly half—assigned this essay at least twice the value given by another teacher. Why should one teacher rate it 10 and another 23? From a study of the results one might say that this question should have been rated 17 to 19, ratings above or below these being probably either too liberal or too low.

The third question, counting 10 points, was a question in syntax. A selection was given and the syntax of certain clauses and phrases was asked. The student omitted certain parts of his answer so that the maximum attainable was 6. Six was his original rating. The Regents rated this question 5, as did twenty-seven of the teachers, three others rating it 4; and one, $3\frac{1}{2}$. This question showed the greatest uniformity; and naturally, because it was possible to form a definite uniform judgment concerning a grammatical fact. This was not so feasible in the essay or the letter.

The fourth question consisted of a selection describing certain traits of Shakespeare's character. These four questions were to be answered: (*a*) Mention in one word the single trait of Shakespeare's character that is emphasized in this paragraph (2 counts). (*b*) Mention *two* other qualities which, according to the paragraph

Shakespeare possessed (4). (*c*) What similar impression is left by the Waverley novels and by Shakespeare's work (3). (*d*) What quality of Shakespeare is showing by his taking offense at Greene's attack and by the Sir Thomas Lucy incident (3).

The ratings given show a wide range, running from 5 to 12. Table IV indicates this variability.

TABLE IV

Rating	No. of Teachers	Rating	No. of Teachers
5	4	9	5
6	6	10	2
7	6	11	2
8	5	12	1

As was seen in the other illustrations, the ratings cover a wide range, four teachers giving the answer a value of 5 and five others thinking it worth 10 or better. What causes such a diversity of judgments? It was rated originally 6, rated 5 when the original examiner re-rated it, and rated 3 by the Regents (unquestionably too low). The average rating given by the thirty-one teachers was 7.7, the median being 7; 50 per cent of the rating being between $6\frac{1}{2}$ and $7\frac{1}{2}$. These latter limits might be regarded as probably just ratings, those above $7\frac{1}{2}$ too generous, and those below $6\frac{1}{2}$ too severe.

In the fifth question the candidate wrote a paragraph of about 100 words on why *Silas Marner* (one of a wide choice) was a book worth studying. He was permitted to argue for or against the study of his book. The variations in the ratings given approximate very closely the normal distribution and the normal probability curve. The maximum allowed was 10. The accompanying table indicates the distribution.

TABLE V

Rating	No. of Teachers	Rating	No. of Teachers
4	2	7	12
5	4	8	2
6	10	9	1

Again we meet this same striking diversity of judgment; two teachers rate his argument as worth 4 out of 10, and three

others rate it as being worth at least 8 out of 10—twice as much as the first group. Why? Trained teachers of English ought to be able to tell whether or not an argumentative discourse was presented adequately or not; and if they are not entirely uniform in their judgments, they ought to approximate uniformity a little more closely than seems to be indicated here.

In the sixth question this student wrote a paragraph of about 150 words giving his interpretation of the conception of liberty indicated in Burke's "Speech on Conciliation with America." (Credit allowed 20.) The greatest variation appears in the judgment of this answer, the ratings ranging from 2 to 17. Table VI and Figure 2 display the results.

TABLE VI

Rating	No. of Teachers	Rating	No. of Teachers
2	1	12	3
5	2	13	4
8	1	14	2
10	12	15	3
11	2	17	1

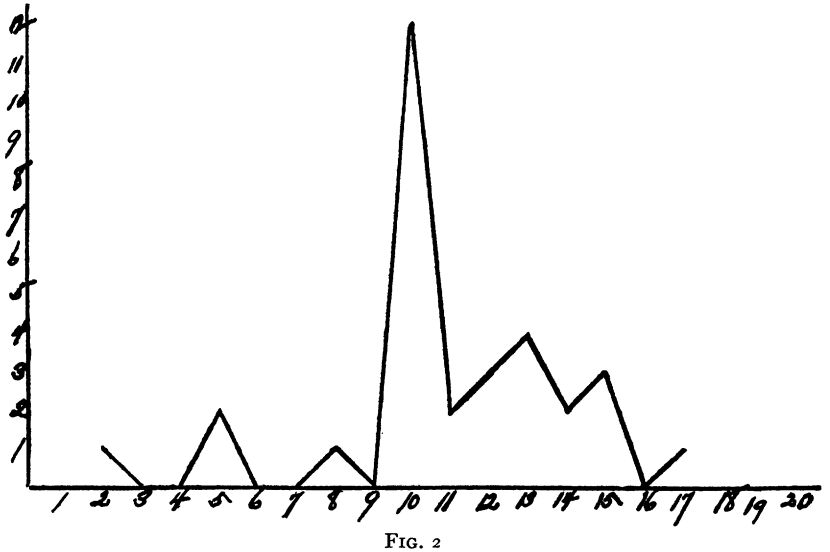

FIG. 2

From this table it is evident that four teachers thought it was worth from 3 to $8\frac{1}{2}$ times as much as did three other teachers. On

what grounds? What standards of judgment were used in this? It is evident that these must have been very vague. A standard of any type would have indicated this question as worth much more than 2, and certainly much less than 17. Probably a rating of 10, 11, and 12 might have been regarded as a good valuation.

The seventh and last question called upon the pupil to quote ten consecutive lines of poetry from memory and interpret this passage quoted, the quotation receiving 5 counts and the interpretation 3 counts. The answer contained *nine* lines of Antony's address at Caesar's burial incorrectly quoted. This, however, did not prevent one of the examiners from rating it 8 out of a possible 8. In fact he received every possible rating from zero, as Table VII and Figure 3 distribution show.

In the preceding discussion I have tried to withstand the temptation to give the foregoing figures a too refined statistical treatment. I am interested in statistical interpretation rather than in technique. What do the foregoing figures tell us concerning the probable worth of the paper? What do they indicate concerning the abilities of the teachers as judges of a paper? Since the answer to the first question depends to some extent upon the second, let us try to answer the second question first.

In one question only, the syntax question, was there any uniformity, twenty-seven of the thirty-one teachers giving the same rating. In every other question there was a very wide range of ratings; as, for instance, every possible rating in the last question, ratings anywhere from 2 to 17 out of a possible 20 in the sixth question, etc. In a question that is taught by the scientific method —as in syntax—uniformity is not a difficult matter to obtain. But in a question that calls for literary appreciation, or criticism or original composition, the evaluation is based upon that variable quantity, "the teacher's judgment," rather than upon scientific laws and principles. Laws and principles for judging literature are not so exactly defined as are the laws of grammar; and for that reason uniform judgments are not readily obtainable.

Another factor that must not be overlooked is the peculiar make-up or composition of the high-school teacher of English. English divides itself logically into a number of subdivisions such

as literature, comparative or dramatic; prose; poetry; grammar; rhetoric; composition, etc. Our English teachers, because of their university training and inclinations, are usually specialists. One teacher may be particularly interested in poetry, another in dramatic literature; but no matter what one's aptitudes may be,

TABLE VII

Rating	No. of Teachers	Rating	No. of Teachers
1	1	5	2
2	11	6	3
3	3	7	1
4	9	8	1

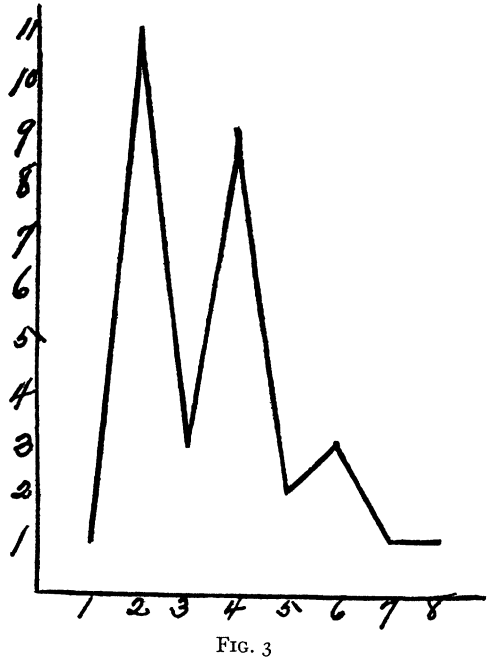

FIG. 3

the teacher teaches all the various subdivisions under the general head of English. These natural prejudices are brought to bear in marking answer papers written in examinations. A paper which contains a false syntax falls into the hands of a grammarian, the examiner gnashes his teeth and dire consequences follow. On the other hand let us suppose this same paper falls into the hands of a

poet or one interested in poetry. Suppose the paper contained a beautiful thought clearly expressed. What if it does contain a misspelled word or a grammatical fault? The examiner knows that the pupil does or should know how to spell that word; and with a glow of satisfaction he gives a high rating.

What are we going to do about it? To diagnose the case is easy enough; but how about therapy? Can we prescribe a specific? It seems to me that right here is where education has been making the same mistake that the medical profession has made. We are stressing "diagnosis" too much. Tests and scales are elaborately prepared and sent broadcast because of their "diagnostic" value. Suppose after applying these "tests" I do discover that a boy cannot "factor" in algebra, how will their tests help to make that boy *factor?* Physicians are hailed because they recognized (diagnosed) some rare disease, but the fact is only too frequently overlooked that after the recognition they are not able to do any more for their patient than was done twenty-five, fifty, or a hundred years ago. Diagnosis has made wonderful strides, but therapy lags very far behind; and the educational world is imitating the medical world and repeating its errors.

Some suggestions at this time by way of an attempted cure might not be out of place. One essential element that ought to be decided upon is the relative value of the form and the content of an answer. How much is a good thought worth? How bad is an error in spelling or grammar as contrasted with a slovenly thought-out idea? A beginning made here might tend to do a little toward answering the question as to how good or how bad a pupil's work is when taken as a whole.

An attempt was made to find out whether a teacher was a fair marker, a good marker, an easy marker, a severe marker, or an erratic marker as compared with the department as a whole. The method was as follows: In question 1, 50 per cent of the teachers on either side of the median rating 11 gave this question ratings of 10, 11, and 12. These, therefore, seemed to be the fair ratings for this question. Opposite each of these teachers' names was placed a check (✓). Such teachers as rated the question more than 12 were marked with a plus (+) sign; they were easy markers. Such

teachers as rated the question less than 10 were marked with a minus (−) sign; they were severe markers. The same procedure was followed with the other questions. The results were as follows:

TABLE VIII

Teacher	I	II	III	IV	V	VI	VII	
A	✓	+	✓	+	✓	✓	✓	
B	✓	✓	✓	+	✓	✓	−	
C	+	✓	−	+	✓	✓	+	**
D	✓	✓	−	✓	+	+	✓	
E	✓	+	✓	✓	−	✓	✓	
F	−	+	✓	+	✓	+	−	
G	✓	+	✓	+	✓	+	✓	***
H	✓	✓	−	−	✓	+	−	
I	✓	−	✓	✓	✓	−	−	*
J	✓	−	✓	−	✓	−	−	*
K	✓	✓	✓	−	✓	✓	✓	
L	✓	✓	✓	✓	−	✓	+	
M	✓	+	✓	+	−	✓	−	**
N	+	✓	✓	+	✓	−	+	
O	−	+	✓	✓	−	✓	✓	
P	+	+	✓	✓	✓	✓	+	***
Q	✓	+	✓	✓	✓	✓	−	
R	✓	+	✓	✓	✓	✓	✓	
S	✓	−	✓	−	✓	−	−	*
T	✓	−	✓	+	+	✓	+	
U	✓	−	✓	+	−	✓	+	**
V	+	✓	✓	+	✓	✓	+	
X	✓	+	✓	+	✓	+	−	
Y	✓	✓	−	+	✓	✓	−	
Z	✓	+	✓	✓	✓	+	✓	
A1	✓	−	✓	+	+	+	✓	
B1	✓	✓	✓	✓	✓	+	−	
C1	✓	+	✓	+	✓	+	−	
D1	+	+	✓	✓	✓	✓	✓	
E1	+	✓	✓	+	✓	✓	✓	
F1	+	✓	✓	✓	✓	+	+	***

From the foregoing a teacher may be judged easy, hard, fair (good), or erratic, by the number of checks, plus signs, or minus signs found against his or her score. Only one teacher (R) had a score that might be regarded as fair as judged by the teachers as a whole. How to judge all the teachers from this table might be rather difficult. However, the extreme cases clearly define themselves. Teachers I, J, and S were unquestionably hard markers as is evidenced by the many minus signs in their scores. Teachers G, P, and F1 were easy markers; they had many plus signs.

Teachers C and M were rather erratic, marking some questions fairly, some severely, and some leniently.

What was the paper worth? This paper was marked in committee last June. Had it passed through that committee of seven teachers who had given the seven questions their lowest ratings, the paper would have received a mark of 32 per cent. If on the other hand it had received a marking at the hands of those who had been most generous in this experiment it would have been rated 89 per cent. It had been marked originally 73 per cent; Albany evaluated it 59 per cent. By taking the judgments of the teachers as a whole and evaluating a rating for each question the probable value of the paper is found to be 61½ per cent. The following table gives the facts:

TABLE IX

	I	II	III	IV	V	VI	VII	Total
Lowest	7	10	3½	5	4	2	1	32½
Highest	15	23	5	12	9	17	8	89
Original	13	20	6	6	8	15	5	73
Regents	10	18	5	3	6	13	4	59
Probable	11	18	5	7	6½	11	3	61½

Can it be made possible to place upon a scientific basis the evaluation of a piece of work done by a pupil, either in regular recitation or in examination? Probably only after a number of comparative experiments, similar to this one, have been performed.